# Data and Replication files for

**Quantiles of the gain distribution of an early childhood intervention (replication data)**, by Erich Battistin, Carlos Lamarche, and Enrico Rettore. Journal of Applied Econometrics, forthcoming. Last revision: April 29, 2024.

## A) Data

Data from the Infant Health and Development Program (IHDP) is accessible under specific terms and conditions via the following link:

https://www.icpsr.umich.edu/web/ICPSR/studies/9795

This site provides several files in ASCII format containing the raw data. The Stata codes below reference specific files and variables from this collection to compile the working sample. Replication of all results is conditional on data access.

## B) Replication of results:

All steps below can be replicated from "${foldwork}\shell.do". Stata, version 15

Output files are saved into:

cap global foldgr  "$foldwork\graphs"

cap global foldlog  "$foldwork\output"

## Step 1: Derivation of the working sample

"${foldwork}\import_data.do"

"${foldwork}\make_data.do"

"${foldwork}\macros-setup.do"

The working sample is the IHDP Primary Analysis Group (N=985)

We use the following measures for the health endowment, recorded at birth:

- NHIBW: Health Index - see page 279 of the IHDP codebook file; it's a residual not explained by birth weight, using duration from date of birth to date of hospital discharge
- INV_BW: Weight
- F3AV4: Length
- F3AV5: Head Circumference
- F3AMAGE: Maternal Age
- INV_ANGA: Gestational Age

We keep only infants without missing data on the variables above. **The final sample size is N=929. The working file is "${folddata}\datapaper.dta"**.

We derive the following demographics at birth:

- Infant was a boy
- Infant is a first-born
- Age of the mother
- Mother was married
- Mother was high school dropout
- Mother was high school graduate
- Mother was college graduate
- Black Maternal Race
- Hispanic Maternal Race

We use the following outcomes after 36 months:

- F57_1_ Cognitive Development Index
- THCS_CUM Morbidity Index
- GHRI_36 General Health Ratings Index

## Step 2: Balancing tests and descriptives

"${foldwork}\macros-setup.do"

"${foldwork}\descriptives.do"

Table 1 is obtained first, then Table 2. **Results were copied manually to create the latex tables in the text**.

## Step 3: Quantile treatment effects

"${foldwork}\macros-setup.do"

"${foldwork}\rank-invariance-rev-sep-2023.do"

**Figure 1 is obtained first using 100 bootstrap samples**. Bootstrap estimates are saved in files "${foldlog}\QTE_`v'.dta", which are then used to plot results. The lines in the code running bootstrap are commented out.

**The rank similarity test in in Frandsen and Lefgren (2018) is obtained next**.

## Step 4: Measurement equations

"${foldwork}\macros-setup.do"

"${foldwork}\IHDP-measurements-equations-june-2023.do"

**The file starts by showing results from a PCA** using the four anthropometric measurements:

macro def W_c INV_BW  // birth weight

macro def W_d INV_ANGA  // gestational age

macro def W_e F3AV5  // head circumference at birth (cm)

macro def W_f F3AV4  // infant length at birth (cm)

**It also computes differences of the first PC across birth weight groups**.

All variables above and the outcomes:

- F57_1_  Cognitive Development Index
- THCS_CUM  Morbidity Index
- GHRI_36  General Health Ratings Index

are "residualized" using regressions on:

label var F3AMAGE  "Age of the mother"

label var married  "Mother was married"

label var highdrop  "Mother was high school dropout"

label var highgrad  "Mother was high school graduate"

label var collgrad  "Mother was college graduate"

label var BLACK  "Black Maternal Race"

label var HISPANIC  "Hispanic Maternal Race"

separately for treatment (treat=1) and control (treat=0) infants.

We then standardize residuals to have zero mean and unit variance in the sample.

**Appendix**: Factor loadings are estimated from 2SLS regressions. The same estimates can be used to obtain measurement error variances, although we do not present this in the paper. Measurement error variances are obtained by combining equations using sureg.

**Main text**: measurement error variances are estimated using nlsur.

## Step 5: Computation of QCD

The files:

do "${foldwork}\simex-SIMstep.do"

do "${foldwork}\simex-EXstep.do"

implement the simulation and extrapolation steps of SIMEX, respectively.

The file:

do "${foldwork}\ simex-output.do"

obtains estimates of QCD, and saves them in the files:

saveold "${foldlog}\simex-data-std.dta", replace // these are point estimates

saveold "${foldlog}\simex-data-std-boot`i'.dta", replace // these are bootstrap estimates

Estimates of QCD obtained from raw data, that is, using birth weight instead of the latent factor H, are obtained in:

"${foldwork}\raw-data-output.do"

Estimates of loadings from the location-shift specification are obtained in:

"${foldwork}\ simex-graphs-bootstrap-july23.do"

## Step 6: Computation of QCD by income group

This block of files is organized as those in Step 5. Point estimates are saved in:

saveold "${foldlog}\simex-data-std-high.dta", replace

saveold "${foldlog}\simex-data-std-high-boot`i'.dta", replace

saveold "${foldlog}\simex-data-std-low.dta", replace

saveold "${foldlog}\simex-data-std-low-boot`i'.dta", replace